CLAIMS

What is claimed is:

1. A method for creating a description of a document of a remote network data source for later identification of the document, comprising:

(a) receiving information from a user about a document on a remote network data site;

(b) creating a document identifier based on the user-input information, wherein the document identifier identifies the particular document;

(c) retrieving a markup language description defining properties of elements of a document in a markup language;

(d) analyzing the document and the content of the document utilizing the document identifier and the markup language description;

(e) generating a description of the document based on the analysis; and

(f) storing the document description.

2. The method as recited in claim 1, wherein information received from the user includes at least one of: an identification of content of interest in the document, guidelines for recognizing a document, and guidelines for recognizing content elements of interest.

3. The method as recited in claim 1, wherein the document description contains a list of elements of interest and element properties for the elements of interest.

4. The method as recited in claim 1, wherein the analysis of the content is for identifying elements of interest of the content of the document.

5. The method as recited in claim 4, wherein the markup language description is used to identify properties of each of the elements of interest.

6. The method as recited in claim 5, wherein the elements of interest of the content are identified based on properties of each element.

7. The method as recited in claim 1, wherein the document analysis includes comparing the document to at least one other document, wherein the document description is modified to reflect at least one difference between the documents.

8. The method as recited in claim 1, further comprising comparing the document to at least one other document, wherein document descriptions of each of the documents are modified to reflect at least one difference between the documents.

9. The method as recited in claim 1, wherein the document is modified, wherein the document identifier is modified, wherein the modified document is analyzed for modifying the document description.

10. The method as recited in claim 9, wherein the document analysis includes comparing the modified document to at least one other document, wherein the document description is modified to reflect at least one difference between the documents.

11. The method as recited in claim 1, wherein the method is performed during creation of a transaction pattern.

12. A computer program product for creating a description of a document of a remote network data source for later identification of the document, comprising:

(a) computer code for receiving information from a user about a document on a remote network data site;

(b) computer code for creating a document identifier based on the user-input information, wherein the document identifier identifies the particular document;

(c) computer code for retrieving a markup language description defining properties of elements of a document in a markup language;

(d) computer code for analyzing the document and the content of the document utilizing the document identifier and the markup language description;

(e) computer code for generating a description of the document based on the analysis; and

(f) computer code for storing the document description.

13. The computer program product as recited in claim 12, wherein information received from the user includes at least one of: an identification of content of interest in the document, guidelines for recognizing a document, and guidelines for recognizing content elements of interest.

14. The computer program product as recited in claim 12, wherein the document description contains a list of elements of interest and element properties for the elements of interest.

15. The computer program product as recited in claim 12, wherein the analysis of the content is for identifying elements of interest of the content of the document.

16. The computer program product as recited in claim 12, wherein the document analysis includes comparing the document to at least one other document,

wherein the document description is modified to reflect at least one difference between the documents.

17. The computer program product as recited in claim 12, further comprising computer code for comparing the document to at least one other document, wherein document descriptions of each of the documents are modified to reflect at least one difference between the documents.

18. The computer program product as recited in claim 12, wherein the computer program is executed during creation of a transaction pattern.

19. A system for creating a description of a document of a remote network data source for later identification of the document, comprising:
(a) logic for receiving information from a user about a document on a remote network data site;
(b) logic for creating a document identifier based on the user-input information, wherein the document identifier identifies the particular document;
(c) logic for retrieving a markup language description defining properties of elements of a document in a markup language;
(d) logic for analyzing the document and the content of the document utilizing the document identifier and the markup language description;
(e) logic for generating a description of the document based on the analysis; and
(f) logic for storing the document description.

20. A method for creating a description of content of a remote network data source for later identification of the content, comprising:
(a) receiving information from a user about content on a remote network data site;
(b) creating a content identifier based on the user-input information, wherein the content identifier identifies the particular content;

(c) retrieving a markup language description defining properties of elements of the content in a markup language;

(d) analyzing the content utilizing the content identifier and the markup language description;

(e) generating a description of the content based on the analysis; and

(f) storing the content description.

21. The method as recited in claim 20, wherein information received from the user includes at least one of: an identification of content elements of interest, guidelines for recognizing content, and guidelines for recognizing content elements of interest.

22. The method as recited in claim 20, wherein the content description contains a list of elements of interest and element properties for the elements of interest.

23. The method as recited in claim 20, wherein the content is a document.

24. The method as recited in claim 23, wherein a description of content items of the document is stored.

25. A method for identifying a document, comprising:

(a) receiving a document;

(b) receiving document descriptions of several documents;

(c) comparing the document descriptions with the document;

(d) calculating a document recognition score for each of the document descriptions based on a likelihood that the document description matches the document;

(e) selecting a document description based at least in part on the document recognition scores; and

(f) identifying the document based on the selected document description.

26. The method as recited in claim 25, wherein the document recognition score is based at least in part on recognizing properties of elements of the documents in the document descriptions.

27. The method as recited in claim 26, wherein each of the properties is given a weight.

28. The method as recited in claim 27, wherein the weights are normalized.

29. The method as recited in claim 28, wherein selected elements of the document are each given a content recognition score, wherein the content recognition score is a weighted sum of values returned by a property evaluation function weighted with the normalized weight of the property, wherein the content recognition scores are used to determine whether each content element is present.

30. The method as recited in claim 29, wherein the document recognition score for each document description is calculated using the formula $S_k = \sum_{i=1}^{N} p_i R_i$, wherein N is a number of elements of interest in the document, p_i is the presence weight of element I, and R_i is a function of the content recognition score for element i.

31. The method as recited in claim 25, wherein the selection of the document is based on the document recognition scores and deviation, wherein the deviation is computed from the document recognition scores.

32. The method as recited in claim 31, wherein a document description with a high document recognition score relative to other candidate document descriptions and a deviation above a predetermined threshold is selected.

33. The method as recited in claim 31, wherein a document description with a low document recognition score relative to other candidate document descriptions and a deviation above a predetermined threshold is selected.

34. The method as recited in claim 31, wherein the deviation is calculated using the formula $d_{recognition} = \left(\sum_{i=1}^{k-1} \frac{1}{|S_i - S_k|} + \sum_{i=k+1}^{T} \frac{1}{|S_i - S_k|} \right)^{-1}$, where S_i is the recognition score for document i, k is the index of the matched document, and T is the number of candidate documents.

35. The method as recited in claim 25, further comprising pruning for reducing processing.

36. The method as recited in claim 25, further comprising retrieving portions of the document.

37. The method as recited in claim 36, wherein the portion is retrieved using a content identifier pre-associated with the portion.

38. The method as recited in claim 25, wherein the method is performed during replay of a transaction pattern.

39. The method as recited in claim 25, wherein a hint is received, wherein the hint indicates that one document description is more likely to match the document than another document description.

40. The method as recited in claim 38, wherein the hint includes an order of processing by which one document description is processed in respect to other documents descriptions.

41. The method as recited in claim 38, wherein the hint includes a hint threshold, wherein the hint threshold is a value for determining when a document description matches the document.

42. The method as recited in claim 38, wherein the hint includes an order of processing by which one document description is processed in respect to other documents descriptions, and a hint threshold, wherein the hint threshold is a value that tells the algorithm when the document is matched.

43. A computer program product for identifying a document, comprising:
(a) computer code for receiving a document;
(b) computer code for receiving document descriptions of several documents;
(c) computer code for comparing the document descriptions with the document;
(d) computer code for calculating a document recognition score for each of the document descriptions based on a likelihood that the document description matches the document;
(e) computer code for selecting a document description based at least in part on the document recognition scores; and
(f) computer code for identifying the document based on the selected document description.

44. A method for identifying content, comprising:
(a) receiving several content elements;
(b) receiving a content description of a desired content element;
(c) comparing the content description with the received content elements;
(d) calculating a content recognition score for each of the content elements based on a likelihood that the content description matches the content element; and

(e) selecting a matching content based at least in part on the content recognition scores.

45. A method for creating a description of a document of a remote network data source for later identification of the document, comprising:

(a) receiving information from a user about a document on a remote network data site, wherein the information received from the user includes at least one of: an identification of content of interest in the document, guidelines for recognizing a document, and guidelines for recognizing content elements of interest;

(b) creating a document identifier based on the user-input information, wherein the document identifier identifies the particular document;

(c) retrieving a markup language description defining properties of elements of a document in a markup language;

(d) comparing the document to at least one other document utilizing the document identifier and the markup language description;

(e) analyzing the content of the document utilizing the document identifier and the markup language description for identifying elements of interest of the content of the document;

(f) generating a description of the document based on the comparison and analysis, wherein the document description contains a list of the elements of interest and element properties for the elements of interest, wherein the document description reflects at least one difference between the document and the at least one other document; and

(g) storing the document description.